

# Algebraic models for higher-order correlations

Lek-Heng Lim and Jason Morton

U.C. Berkeley and Stanford Univ.

December 15, 2008

# Tensors as hypermatrices

Up to choice of bases on  $U, V, W$ , a tensor  $A \in U \otimes V \otimes W$  may be represented as a hypermatrix

$$\mathcal{A} = \llbracket a_{ijk} \rrbracket_{i,j,k=1}^{l,m,n} \in \mathbb{R}^{l \times m \times n}$$

where  $\dim(U) = l, \dim(V) = m, \dim(W) = n$  if

- 1 we give it coordinates;
- 2 we ignore covariance and contravariance.

Henceforth, tensor = hypermatrix.

## Multilinear matrix multiplication

- Matrices can be multiplied on left and right:  $A \in \mathbb{R}^{m \times n}$ ,  $X \in \mathbb{R}^{p \times m}$ ,  $Y \in \mathbb{R}^{q \times n}$ ,

$$C = (X, Y) \cdot A = XAY^T \in \mathbb{R}^{p \times q},$$

$$c_{\alpha\beta} = \sum_{i,j=1}^{m,n} x_{\alpha i} y_{\beta j} a_{ij}.$$

- 3-tensors can be multiplied on three sides:  $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$ ,  $X \in \mathbb{R}^{p \times l}$ ,  $Y \in \mathbb{R}^{q \times m}$ ,  $Z \in \mathbb{R}^{r \times n}$ ,

$$C = (X, Y, Z) \cdot \mathcal{A} \in \mathbb{R}^{p \times q \times r},$$

$$c_{\alpha\beta\gamma} = \sum_{i,j,k=1}^{l,m,n} x_{\alpha i} y_{\beta j} z_{\gamma k} a_{ijk}.$$

- Correspond to change-of-bases transformations for tensors.
- Define 'right' (covariant) multiplication by  $(X, Y, Z) \cdot \mathcal{A} = \mathcal{A} \cdot (X^T, Y^T, Z^T)$ .

# Symmetric tensors

- Cubical tensor  $\llbracket a_{ijk} \rrbracket \in \mathbb{R}^{n \times n \times n}$  is **symmetric** if

$$a_{ijk} = a_{ikj} = a_{jik} = a_{jki} = a_{kij} = a_{kji}.$$

- For order  $p$ , invariant under all permutations  $\sigma \in \mathfrak{S}_p$  on indices.
- $S^p(\mathbb{R}^n)$  denotes set of all order- $p$  symmetric tensors.
- Symmetric multilinear matrix multiplication  $\mathcal{C} = (X, X, X) \cdot \mathcal{A}$  where

$$c_{\alpha\beta\gamma} = \sum_{i,j,k=1}^{l,m,n} x_{\alpha i} x_{\beta j} x_{\gamma k} a_{ijk}.$$

# Examples of symmetric tensors

- Higher order derivatives of real-valued multivariate functions.
- Moments of a vector-valued random variable  $\mathbf{x} = (x_1, \dots, x_n)$ :

$$S_p(\mathbf{x}) = \left[ \left[ E(x_{j_1} x_{j_2} \cdots x_{j_p}) \right]_{j_1, \dots, j_p=1}^n \right]^n.$$

- Cumulants of a random vector  $\mathbf{x} = (x_1, \dots, x_n)$ :

$$\mathcal{K}_p(\mathbf{x}) = \left[ \left[ \sum_{A_1 \sqcup \dots \sqcup A_q = \{j_1, \dots, j_p\}} (-1)^{q-1} (q-1)! E\left(\prod_{j \in A_1} x_j\right) \cdots E\left(\prod_{j \in A_q} x_j\right) \right]_{j_1, \dots, j_p=1}^n \right]^n.$$

# Cumulants

- In terms of log characteristic and cumulant generating functions,

$$\begin{aligned}\kappa_{j_1 \dots j_p}(\mathbf{x}) &= \frac{\partial^p}{\partial t_{j_1} \dots \partial t_{j_p}} \log \mathbf{E}(\exp(\langle \mathbf{t}, \mathbf{x} \rangle)) \Big|_{\mathbf{t}=\mathbf{0}} \\ &= (-1)^p \frac{\partial^p}{\partial t_{j_1} \dots \partial t_{j_p}} \log \mathbf{E}(\exp(i \langle \mathbf{t}, \mathbf{x} \rangle)) \Big|_{\mathbf{t}=\mathbf{0}}.\end{aligned}$$

- In terms of Edgeworth expansion,

$$\log \mathbf{E}(\exp(i \langle \mathbf{t}, \mathbf{x} \rangle)) = \sum_{\alpha=0}^{\infty} i^{|\alpha|} \kappa_{\alpha}(\mathbf{x}) \frac{\mathbf{t}^{\alpha}}{\alpha!}, \quad \log \mathbf{E}(\exp(\langle \mathbf{t}, \mathbf{x} \rangle)) = \sum_{\alpha=0}^{\infty} \kappa_{\alpha}(\mathbf{x}) \frac{\mathbf{t}^{\alpha}}{\alpha!},$$

$\alpha = (\alpha_1, \dots, \alpha_n)$  is a multi-index,  $\mathbf{t}^{\alpha} = t_1^{\alpha_1} \dots t_n^{\alpha_n}$ ,  $\alpha! = \alpha_1! \dots \alpha_n!$ .

- For each  $\mathbf{x}$ ,  $\mathcal{K}_p(\mathbf{x}) = \llbracket \kappa_{j_1 \dots j_p}(\mathbf{x}) \rrbracket \in S^p(\mathbb{R}^n)$  is a symmetric tensor.
- [Fisher, Wishart; 1932]

# Properties of cumulants

**Multilinearity:** If  $\mathbf{x}$  is a  $\mathbb{R}^n$ -valued random variable and  $A \in \mathbb{R}^{m \times n}$

$$\mathcal{K}_p(A\mathbf{x}) = (A, \dots, A) \cdot \mathcal{K}_p(\mathbf{x}).$$

**Additivity:** If  $\mathbf{x}_1, \dots, \mathbf{x}_k$  are mutually independent of  $\mathbf{y}_1, \dots, \mathbf{y}_k$ , then

$$\mathcal{K}_p(\mathbf{x}_1 + \mathbf{y}_1, \dots, \mathbf{x}_k + \mathbf{y}_k) = \mathcal{K}_p(\mathbf{x}_1, \dots, \mathbf{x}_k) + \mathcal{K}_p(\mathbf{y}_1, \dots, \mathbf{y}_k).$$

**Independence:** If  $I$  and  $J$  partition  $\{j_1, \dots, j_p\}$  so that  $\mathbf{x}_I$  and  $\mathbf{x}_J$  are independent, then

$$\kappa_{j_1 \dots j_p}(\mathbf{x}) = 0.$$

**Support:** There are no distributions where

$$\mathcal{K}_p(\mathbf{x}) \begin{cases} \neq 0 & 3 \leq p \leq n, \\ = 0 & p > n. \end{cases}$$

# Examples of cumulants

**Univariate:**  $\mathcal{K}_p(x)$  for  $p = 1, 2, 3, 4$  are mean, variance, skewness, kurtosis (unnormalized)

**Discrete:**  $x \sim \text{Poi}(\lambda)$ ,  $\mathcal{K}_p(x) = \lambda$  for all  $p$ .

**Continuous:**  $x \sim U([0, 1])$ ,  $\mathcal{K}_p(x) = B_p/p$  where  $B_p = p$ th Bernoulli number.

**Nonexistent:**  $x \sim t(3)$ ,  $\mathcal{K}_p(x)$  does not exist for all  $p \geq 3$ .

**Multivariate:**  $\mathcal{K}_1(\mathbf{x}) = E(\mathbf{x})$  and  $\mathcal{K}_2(\mathbf{x}) = \text{Cov}(\mathbf{x})$ .

**Discrete:**  $\mathbf{x} \sim \text{Mult}(n, \mathbf{q})$ ,

$$\kappa_{j_1 \dots j_p}(\mathbf{x}) = n \frac{\partial^p}{\partial t_{j_1} \dots \partial t_{j_p}} \log(q_1 e^{t_1 x_1} + \dots + q_k e^{t_k x_k}) \Big|_{t_1, \dots, t_k = 0}.$$

**Continuous:**  $\mathbf{x} \sim N(\mu, \Sigma)$ ,  $\mathcal{K}_p(\mathbf{x}) = 0$  for all  $p \geq 3$ .

# Estimation of cumulants

- How do we estimate  $\mathcal{K}_p(\mathbf{x})$  given multiple observations of  $\mathbf{x}$ ?
- Central and non-central moments are

$$\hat{m}_n = \frac{1}{n} \sum_t (x_t - \bar{x})^n, \quad \hat{s}_n = \frac{1}{n} \sum_t x_t^n, \quad \text{etc.}$$

- Cumulant estimator  $\hat{\mathcal{K}}_p(\mathbf{x})$  for  $p = 1, 2, 3, 4$  given by

$$\begin{aligned}\hat{\kappa}_i &= \hat{m}_i = \frac{1}{n} \hat{s}_i \\ \hat{\kappa}_{ij} &= \frac{n}{n-1} \hat{m}_{ij} = \frac{1}{n-1} (\hat{s}_{ij} - \frac{1}{n} \hat{s}_i \hat{s}_j) \\ \hat{\kappa}_{ijk} &= \frac{n^2}{(n-1)(n-2)} \hat{m}_{ijk} = \frac{n}{(n-1)(n-2)} [\hat{s}_{ijk} - \frac{1}{n} (\hat{s}_i \hat{s}_{jk} + \hat{s}_j \hat{s}_{ik} + \hat{s}_k \hat{s}_{ij}) + \frac{2}{n^2} \hat{s}_i \hat{s}_j \hat{s}_k] \\ \hat{\kappa}_{ijkl} &= \frac{n^2}{(n-1)(n-2)(n-3)} [(n+1) \hat{m}_{ijkl} - (n-1) (\hat{m}_{ij} \hat{m}_{kl} + \hat{m}_{ik} \hat{m}_{jl} + \hat{m}_{il} \hat{m}_{jk})] \\ &= \frac{n}{(n-1)(n-2)(n-3)} [(n+1) \hat{s}_{ijkl} - \frac{n+1}{n} (\hat{s}_i \hat{s}_{jkl} + \hat{s}_j \hat{s}_{ikl} + \hat{s}_k \hat{s}_{ijl} + \hat{s}_l \hat{s}_{ijk}) \\ &\quad - \frac{n-1}{n} (\hat{s}_{ij} \hat{s}_{kl} + \hat{s}_{ik} \hat{s}_{jl} + \hat{s}_{il} \hat{s}_{jk}) + \hat{s}_i^2 (\hat{s}_{jk} + \hat{s}_{jl} + \hat{s}_{kl}) \\ &\quad + \hat{s}_j^2 (\hat{s}_{ik} + \hat{s}_{il} + \hat{s}_{kl}) + \hat{s}_k^2 (\hat{s}_{ij} + \hat{s}_{il} + \hat{s}_{jl}) + \hat{s}_l^2 (\hat{s}_{ij} + \hat{s}_{ik} + \hat{s}_{jk}) \\ &\quad - \frac{6}{n^2} \hat{s}_i \hat{s}_j \hat{s}_k \hat{s}_l].\end{aligned}$$

## In terms of matrix multiplication

Data often presented as  $Y \in \mathbb{R}^{m \times n}$ , e.g. gene  $\times$  microarray, text  $\times$  document, person  $\times$  image, user  $\times$  movie, webpage  $\times$  webpage etc.

*“And so we now have eigengenes, eigenarrays and eigenexpression in the world of transcriptomics, eigenproteins and eigenprofiles in proteomics, eigenpathways in metabolomics, and eigenSNPs in genetics. There are also eigenimages and eigenfaces in image analysis, and eigenpatterns in seismology. In fact, if you put any word you like after eigen- into a Google query box and hit return, I guarantee a result. Yes, even eigenresult and eigenGoogle!”*

— Terry Speed, IMS Bulletin, April 2008

- Mean centered, otherwise  $\mathbf{y} = \mathbf{x} - E(\mathbf{x})$ .
- $\hat{\mathcal{K}}_1(\mathbf{y}) = \mathbf{0}$ .
- $\hat{\mathcal{K}}_2(\mathbf{y}) = \frac{1}{n-1} \mathbf{Y}\mathbf{Y}^\top = \frac{1}{n-1} (\mathbf{Y}, \mathbf{Y}) \cdot I_{n \times n}$ .
- $\hat{\mathcal{K}}_3(\mathbf{y}) = \frac{n}{(n-1)(n-2)} (\mathbf{Y}, \mathbf{Y}, \mathbf{Y}) \cdot \mathcal{I}_{n \times n \times n}$ .
- $\mathcal{I}_{n \times n \times n} = \llbracket \delta_{ijk} \rrbracket \in S^3(\mathbb{R}^n)$  is the ‘Kronecker delta tensor’, i.e.  $\delta_{ijk} = 1$  if  $i = j = k$  and  $\delta_{ijk} = 0$  otherwise.

# Factor analysis

- Linear generative model

$$\mathbf{y} = A\mathbf{s} + \boldsymbol{\varepsilon}$$

noise  $\boldsymbol{\varepsilon} \in \mathbb{R}^m$ , factor loadings  $A \in \mathbb{R}^{m \times r}$ , hidden factors  $\mathbf{s} \in \mathbb{R}^r$ , observed data  $\mathbf{y} \in \mathbb{R}^m$ .

- Do not know  $A$ ,  $\mathbf{s}$ ,  $\boldsymbol{\varepsilon}$ , but need to recover  $\mathbf{s}$  and sometimes  $A$  from multiple observations of  $\mathbf{y}$ .
- Time series of observations, get matrices  $Y = [\mathbf{y}_1, \dots, \mathbf{y}_n]$ ,  $S = [\mathbf{s}_1, \dots, \mathbf{s}_n]$ ,  $E = [\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n]$ , and

$$Y = AS + E.$$

**Factor analysis:** Recover  $A$  and  $S$  from  $Y$  by a low-rank matrix approximation  $Y \approx AS$

# Principal and independent components analysis

Principal components analysis:  $\mathbf{s}$  Gaussian,

$$\hat{\mathcal{K}}_2(\mathbf{y}) = Q\Lambda_2Q^\top = (Q, Q) \cdot \Lambda_2,$$

$\Lambda_2 \approx \hat{\mathcal{K}}_2(\mathbf{s})$  diagonal matrix,  $Q \in O(n, r)$ , [Pearson; 1901].

Independent components analysis:  $\mathbf{s}$  statistically independent entries,  $\varepsilon$  Gaussian

$$\hat{\mathcal{K}}_p(\mathbf{y}) = (Q, \dots, Q) \cdot \Lambda_p, \quad p = 2, 3, \dots,$$

$\Lambda_p \approx \hat{\mathcal{K}}_p(\mathbf{s})$  diagonal tensor,  $Q \in O(n, r)$ , [Comon; 1994].

What if

- $\mathbf{s}$  not Gaussian, e.g. power-law distributed data in social networks.
- $\mathbf{s}$  not independent, e.g. functional components in neuroimaging.
- $\varepsilon$  not white noise, e.g. idiosyncratic factors in financial modelling.

# Principal cumulant components analysis

- Note that if  $\varepsilon = \mathbf{0}$ , then

$$\mathcal{K}_p(\mathbf{y}) = \mathcal{K}_p(Q\mathbf{s}) = (Q, \dots, Q) \cdot \mathcal{K}_p(\mathbf{s}).$$

- In general, want principal components that account for variation in all cumulants simultaneously

$$\min_{Q \in O(n,r), C_p \in S^p(\mathbb{R}^r)} \sum_{p=1}^{\infty} \alpha_p \|\hat{\mathcal{K}}_p(\mathbf{y}) - (Q, \dots, Q) \cdot C_p\|_F^2,$$

- $C_p \approx \hat{\mathcal{K}}_p(\mathbf{s})$  not necessarily diagonal.
- Appears intractable: optimization over infinite-dimensional manifold

$$O(n, r) \times \prod_{p=1}^{\infty} S^p(\mathbb{R}^r).$$

- Surprising relaxation: optimization over a single Grassmannian  $\text{Gr}(n, r)$  of dimension  $r(n - r)$ ,

$$\max_{Q \in \text{Gr}(n,r)} \sum_{p=1}^{\infty} \alpha_p \|\hat{\mathcal{K}}_p(\mathbf{y}) \cdot (Q, \dots, Q)\|_F^2.$$

- In practice  $\infty = 3$  or  $4$ .

## Some geometric notions

- Set of 'honest' secants of the Veronese variety in  $S^p(\mathbb{R}^n)$  — not closed, not irreducible, difficult to study.
- Symmetric subspace variety in  $S^p(\mathbb{R}^n)$  — closed, irreducible, easy to study.
- Stiefel manifold  $O(n, r)$ : set of  $n \times r$  real matrices with orthonormal columns.  $O(n, n) = O(n)$ , usual orthogonal group.
- Grassmann manifold  $Gr(n, r)$ : set of equivalence classes of  $O(n, r)$  under left multiplication by  $O(n)$ .
- Parametrization of  $S^p(\mathbb{R}^n)$  via

$$Gr(n, r) \times S^p(\mathbb{R}^r) \rightarrow S^p(\mathbb{R}^n).$$

- More generally

$$Gr(n, r) \times \prod_{p=1}^{\infty} S^p(\mathbb{R}^r) \rightarrow \prod_{p=1}^{\infty} S^p(\mathbb{R}^n).$$

## From Stiefel to Grassmann

- Given  $\mathcal{A} \in S^p(\mathbb{R}^n)$ , some  $r \ll n$ , want

$$\min_{X \in O(n,r), \mathcal{C} \in S^p(\mathbb{R}^r)} \|\mathcal{A} - (X, \dots, X) \cdot \mathcal{C}\|_F,$$

- Unlike approximation by secants of Veronese, subspace approximation problem always has a globally optimal solution.
- Equivalent to

$$\max_{X \in O(n,r)} \|(X^\top, \dots, X^\top) \cdot \mathcal{A}\|_F = \max_{X \in O(n,r)} \|\mathcal{A} \cdot (X, \dots, X)\|_F.$$

- Problem defined on a Grassmannian since

$$\|\mathcal{A} \cdot (X, \dots, X)\|_F = \|\mathcal{A} \cdot (XQ, \dots, XQ)\|_F,$$

for any  $Q \in O(r)$ . Only the subspaces spanned by  $X$  matters.

- Equivalent to

$$\max_{X \in \text{Gr}(n,r)} \|\mathcal{A} \cdot (X, \dots, X)\|_F.$$

- Once we have optimal  $X_* \in \text{Gr}(n, r)$ , may obtain  $\mathcal{C}_* \in S^p(\mathbb{R}^r)$  up to  $O(n)$ -equivalence,

$$\mathcal{C}_* = (X_*^\top, \dots, X_*^\top) \cdot \mathcal{A}.$$

## Coordinate-cycling heuristics

- Alternating Least Squares (i.e. Gauss-Seidel) is commonly used for minimizing

$$\Psi(X, Y, Z) = \|\mathcal{A} \cdot (X, Y, Z)\|_F^2$$

for  $\mathcal{A} \in \mathbb{R}^{l \times m \times n}$  cycling between  $X, Y, Z$  and solving a least squares problem at each iteration.

- What if  $\mathcal{A} \in S^3(\mathbb{R}^n)$  and

$$\Phi(X) = \|\mathcal{A} \cdot (X, X, X)\|_F^2?$$

- Present approach: disregard symmetry of  $\mathcal{A}$ , solve  $\Psi(X, Y, Z)$ , set

$$X_* = Y_* = Z_* = (X_* + Y_* + Z_*)/3$$

upon final iteration.

- Better: L-BFGS on Grassmannian.

# Newton/quasi-Newton on a Grassmannian

- Objective  $\Phi : \text{Gr}(n, r) \rightarrow \mathbb{R}$ ,  $\Phi(X) = \|\mathcal{A} \cdot (X, X, X)\|_F^2$ .
- $\mathbf{T}_X$  tangent space at  $X \in \text{Gr}(n, r)$

$$\mathbb{R}^{n \times r} \ni \Delta \in \mathbf{T}_X \quad \iff \quad \Delta^\top X = 0$$

- 1 Compute Grassmann gradient  $\nabla\Phi \in \mathbf{T}_X$ .
- 2 Compute Hessian or update Hessian approximation

$$H : \Delta \in \mathbf{T}_X \rightarrow H\Delta \in \mathbf{T}_X.$$

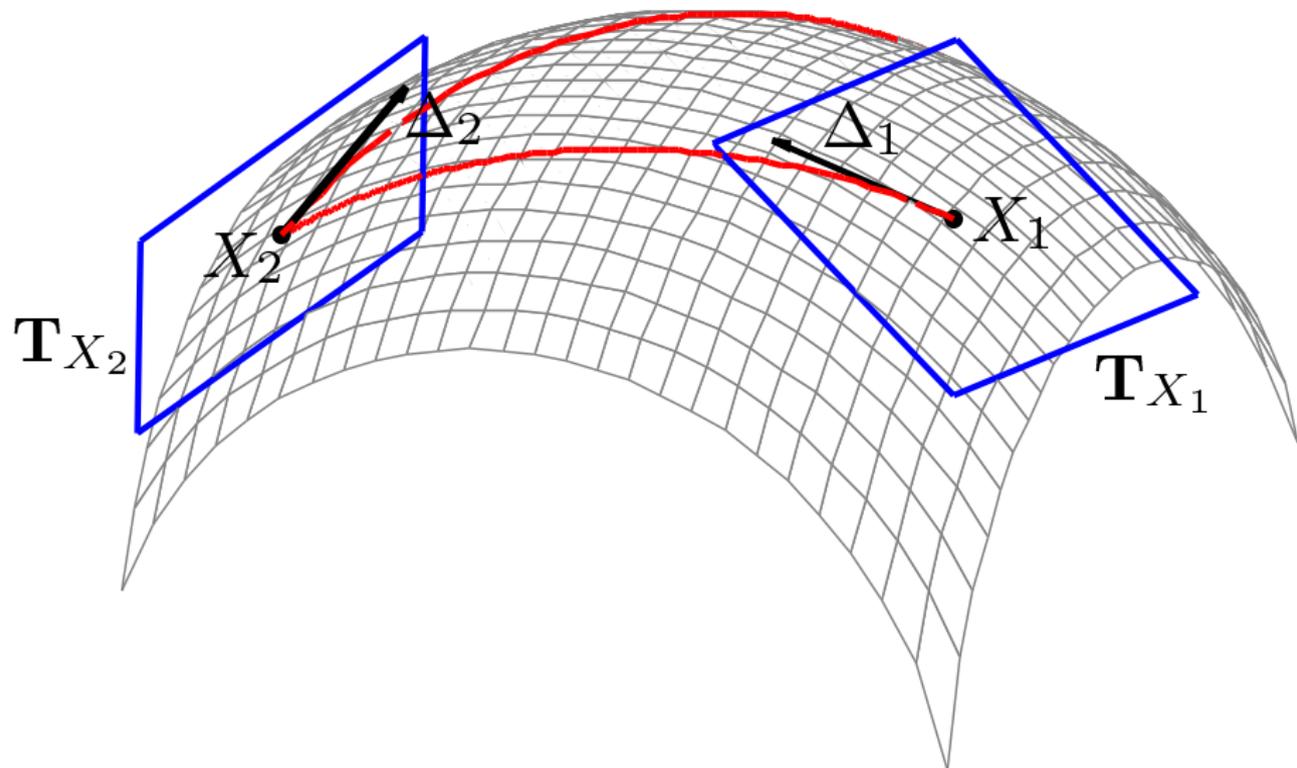
- 3 At  $X \in \text{Gr}(n, r)$ , solve

$$H\Delta = -\nabla\Phi$$

for search direction  $\Delta$ .

- 4 Update iterate  $X$ : Move along geodesic from  $X$  in the direction given by  $\Delta$ .
- [Arias, Edelman, Smith; 1999], [Eldén, Savas; 2008], [Savas, L.; 2008].

# Picture



# BFGS on Grassmannian

The BFGS update

$$H_{k+1} = H_k - \frac{H_k \mathbf{s}_k \mathbf{s}_k^\top H_k}{\mathbf{s}_k^\top H_k \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{y}_k}$$

where

$$\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k = t_k \mathbf{p}_k,$$

$$\mathbf{y}_k = \nabla f_{k+1} - \nabla f_k.$$

On Grassmannian the vectors are defined on different points belonging to different tangent spaces.

## Different ways of parallel transporting vectors

$X \in \text{Gr}(n, r)$ ,  $\Delta_1, \Delta_2 \in \mathbf{T}_X$  and  $X(t)$  geodesic path along  $\Delta_1$

- Parallel transport using global coordinates

$$\Delta_2(t) = T_{\Delta_1}(t)\Delta_2$$

we have also

$$\Delta_1 = X_{\perp} D_1 \quad \text{and} \quad \Delta_2 = X_{\perp} D_2$$

where  $X_{\perp}$  basis for  $\mathbf{T}_X$ . Let  $X(t)_{\perp}$  be basis for  $\mathbf{T}_{X(t)}$ .

- Parallel transport using local coordinates

$$\Delta_2(t) = X(t)_{\perp} D_2.$$

## Parallel transport in local coordinates

All transported tangent vectors have the same coordinate representation in the basis  $X(t)_\perp$  at all points on the path  $X(t)$ .

**Plus:** No need to transport the gradient or the Hessian.

**Minus:** Need to compute  $X(t)_\perp$ .

In global coordinate we compute

- $\mathbf{T}_{k+1} \ni \mathbf{s}_k = t_k T_{\Delta_k}(t_k) \mathbf{p}_k$
- $\mathbf{T}_{k+1} \ni \mathbf{y}_k = \nabla f_{k+1} - T_{\Delta_k}(t_k) \nabla f_k$
- $T_{\Delta_k}(t_k) H_k T_{\Delta_k}^{-1}(t_k) : \mathbf{T}_{k+1} \longrightarrow \mathbf{T}_{k+1}$

$$H_{k+1} = H_k - \frac{H_k \mathbf{s}_k \mathbf{s}_k^\top H_k}{\mathbf{s}_k^\top H_k \mathbf{s}_k} + \frac{\mathbf{y}_k \mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{y}_k}$$

# BFGS

Compact representation of BFGS in Euclidean space:

$$H_k = H_0 + \begin{bmatrix} S_k & H_0 Y_k \end{bmatrix} \begin{bmatrix} R_k^{-\top} (D_k + Y_k^\top H_0 Y_k) R_k^{-1} & -R_k^{-\top} \\ -R_k^{-1} & 0 \end{bmatrix} \begin{bmatrix} S_k^\top \\ Y_k^\top H_0 \end{bmatrix}$$

where

$$S_k = [\mathbf{s}_0, \dots, \mathbf{s}_{k-1}],$$

$$Y_k = [\mathbf{y}_0, \dots, \mathbf{y}_{k-1}],$$

$$D_k = \text{diag} \left[ \mathbf{s}_0^\top \mathbf{y}_0, \dots, \mathbf{s}_{k-1}^\top \mathbf{y}_{k-1} \right],$$

$$R_k = \begin{bmatrix} \mathbf{s}_0^\top \mathbf{y}_0 & \mathbf{s}_0^\top \mathbf{y}_1 & \cdots & \mathbf{s}_0^\top \mathbf{y}_{k-1} \\ 0 & \mathbf{s}_1^\top \mathbf{y}_1 & \cdots & \mathbf{s}_1^\top \mathbf{y}_{k-1} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{s}_{k-1}^\top \mathbf{y}_{k-1} \end{bmatrix}.$$

# L-BFGS

Limited memory BFGS [Byrd et al; 1994]. Replace  $H_0$  by  $\gamma_k I$  and keep the  $m$  most recent  $\mathbf{s}_j$  and  $\mathbf{y}_j$ ,

$$H_k = \gamma_k I + \begin{bmatrix} S_k & \gamma_k Y_k \end{bmatrix} \begin{bmatrix} R_k^{-T} (D_k + \gamma_k Y_k^T Y_k) R_k^{-1} & -R_k^{-T} \\ -R_k^{-1} & 0 \end{bmatrix} \begin{bmatrix} S_k^T \\ \gamma_k Y_k^T \end{bmatrix}$$

where

$$S_k = [\mathbf{s}_{k-m}, \dots, \mathbf{s}_{k-1}],$$

$$Y_k = [\mathbf{y}_{k-m}, \dots, \mathbf{y}_{k-1}],$$

$$D_k = \text{diag} [\mathbf{s}_{k-m}^T \mathbf{y}_{k-m}, \dots, \mathbf{s}_{k-1}^T \mathbf{y}_{k-1}],$$

$$R_k = \begin{bmatrix} \mathbf{s}_{k-m}^T \mathbf{y}_{k-m} & \mathbf{s}_{k-m}^T \mathbf{y}_{k-m+1} & \cdots & \mathbf{s}_{k-m}^T \mathbf{y}_{k-1} \\ 0 & \mathbf{s}_{k-m+1}^T \mathbf{y}_{k-m+1} & \cdots & \mathbf{s}_{k-m+1}^T \mathbf{y}_{k-1} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{s}_{k-1}^T \mathbf{y}_{k-1} \end{bmatrix}.$$

# L-BFGS on the Grassmannian

- In each iteration, parallel transport vectors in  $S_k$  and  $Y_k$  to  $\mathbf{T}_k$ , ie. perform

$$\bar{S}_k = TS_k, \quad \bar{Y}_k = TY_k$$

where  $T$  is the transport matrix.

- No need to modify  $R_k$  or  $D_k$

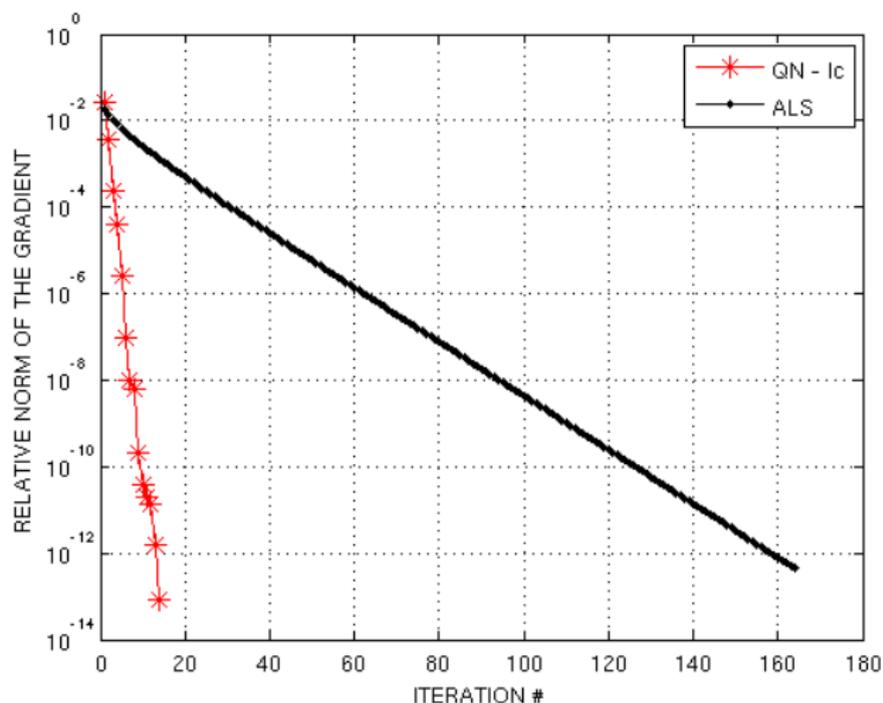
$$\langle \mathbf{u}, \mathbf{v} \rangle = \langle T\mathbf{u}, T\mathbf{v} \rangle$$

where  $\mathbf{u}, \mathbf{v} \in \mathbf{T}_k$  and  $T\mathbf{u}, T\mathbf{v} \in \mathbf{T}_{k+1}$ .

- $H_k$  nonsingular, Hessian is singular. No problem  $\mathbf{T}_k$  at  $\mathbf{x}_k$  is invariant subspace of  $H_k$ , ie. if  $\mathbf{v} \in \mathbf{T}_k$  then  $H_k\mathbf{v} \in \mathbf{T}_k$ .
- [Savas, L.; 2008]

# Convergence

- Compares favorably with Alternating Least Squares.



## Higher order eigenfaces

Principal cumulant subspaces supplement varimax subspace from PCA. Take face recognition for example, **eigenfaces** ( $p = 2$ ) becomes **skewfaces** ( $p = 3$ ) and **kurtofaces** ( $p = 4$ ).

- Eigenfaces: given image  $\times$  pixel matrix  $A \in \mathbb{R}^{m \times n}$  with centered columns where  $m \ll n$ .
- Eigenvectors of pixel  $\times$  pixel covariance matrix  $\mathcal{K}_2^{\text{pixel}} \in S^2(\mathbb{R}^n)$  are the eigenfaces.
- For efficiency, compute image  $\times$  image covariance matrix  $\mathcal{K}_2^{\text{image}} \in S^2(\mathbb{R}^m)$  instead.
- SVD  $A = U\Sigma V^T$  gives both implicitly,

$$\begin{aligned}\mathcal{K}_2^{\text{image}} &= \frac{1}{n}(A^T, A^T) \cdot I_{m \times m} = \frac{1}{n}A^T A = \frac{1}{n}V\Lambda V^T, \\ \mathcal{K}_2^{\text{pixel}} &= \frac{1}{n}(A, A) \cdot I_{n \times n} = \frac{1}{m}AA^T = \frac{1}{m}U\Lambda U^T.\end{aligned}$$

- Orthonormal columns of  $U$ , eigenvectors of  $n\mathcal{K}_2^{\text{pixel}}$ , are the eigenfaces.

# Computing image and pixel skewness

- Want to implicitly compute  $\mathcal{K}_3^{\text{pixel}} \in S^3(\mathbb{R}^n)$ , third cumulant tensor of the pixels (huge).
- Just need projector  $\Pi$  onto the subspace of skewfaces that best explain  $\mathcal{K}_3^{\text{pixel}}$ .
- Let  $A = U\Sigma V^\top$ ,  $U \in O(n, m)$ ,  $\Sigma \in \mathbb{R}^{m \times m}$ ,  $V \in O(m)$ .

$$\begin{aligned}\mathcal{K}_3^{\text{pixel}} &= \frac{1}{m}(A, A, A) \cdot \mathcal{I}_{m \times m \times m} \\ &= \frac{1}{m}(U, U, U) \cdot (\Sigma, \Sigma, \Sigma) \cdot (V^\top, V^\top, V^\top) \cdot \mathcal{I}_{m \times m \times m} \\ \mathcal{K}_3^{\text{image}} &= \frac{1}{n}(A^\top, A^\top, A^\top) \cdot \mathcal{I}_{n \times n \times n} \\ &= \frac{1}{n}(V, V, V) \cdot (\Sigma, \Sigma, \Sigma) \cdot (U^\top, U^\top, U^\top) \cdot \mathcal{I}_{n \times n \times n}\end{aligned}$$

- $\mathcal{I}_{n \times n \times n} = \llbracket \delta_{ijk} \rrbracket \in S^3(\mathbb{R}^n)$  is the 'Kronecker delta tensor', i.e.  $\delta_{ijk} = 1$  iff  $i = j = k$  and  $\delta_{ijk} = 0$  otherwise.

# Computing skewmax projection

- Define  $\mathcal{A} \in S^3(\mathbb{R}^m)$  by

$$\mathcal{A} = (\Sigma, \Sigma, \Sigma) \cdot (V^\top, V^\top, V^\top) \cdot \mathcal{I}_{m \times m \times m}$$

- Want  $Q \in O(m, s)$  and core tensor  $\mathcal{C} \in S^3(\mathbb{R}^s)$  not necessarily diagonal, so that  $\mathcal{A} \approx (Q, Q, Q) \cdot \mathcal{C}$  and thus

$$\mathcal{K}_3^{\text{pixel}} \approx \frac{1}{m}(U, U, U) \cdot (Q, Q, Q) \cdot \mathcal{C} = \frac{1}{m}(UQ, UQ, UQ) \cdot \mathcal{C}.$$

- Solve

$$\min_{Q \in O(m, s), \mathcal{C} \in S^3(\mathbb{R}^s)} \|\mathcal{A} - (Q, Q, Q) \cdot \mathcal{C}\|_F$$

- $\Pi = UQ \in O(n, s)$  is our orthonormal-column projection matrix onto the 'skewmax' subspace.
- Caveat:  $Q$  only determined up to  $O(s)$ -equivalence. Not a problem if we are just interested in the associated subspace or its projector.

# Combining eigen-, skew-, and kurtofaces

Combine information from multiple cumulants:

- Same procedure for the kurtosis tensor (a little more complicated).
- Say we keep the first  $r$  eigenfaces (columns of  $U$ ),  $s$  skewfaces, and  $t$  kurtofaces. Their span is our optimal subspace.
- These three subspaces may overlap; orthogonalize the resulting  $r + s + t$  column vectors to get a final projector.

This gives an orthonormal projector basis  $W$  for the column space of  $A$ ; its

- first  $r$  vectors best explain the pixel covariance  $\mathcal{K}_2^{\text{pixel}} \in S^2(\mathbb{R}^n)$ ,
- next  $s$  vectors, with  $W_{1:r}$ , best explain the pixel skewness  $\mathcal{K}_3^{\text{pixel}} \in S^3(\mathbb{R}^n)$ ,
- last  $t$  vectors, with  $W_{1:r+s}$ , best explain pixel kurtosis  $\mathcal{K}_4^{\text{pixel}} \in S^4(\mathbb{R}^n)$ .

# Advertisement and acknowledgement

Jason Morton, “Algebraic models for multilinear dependence,” SAMSI Workshop on *Algebraic Statistical Models*, Research Triangle Park, NC, January 15–17, 2009.

Thanks:

- J.M. Landsberg
- Berkant Savas
- Henry Wynn